ヒストグラムにおけるAnchor Positionの選択法

金沢大学人間社会研究域経済学経営学系　寒河江　雅彦
㈱日立製作所　山本　敏寛

要　旨　与えられたデータからヒストグラムを作ることは，統計の知識の有無を問わずにデータの大まかな特徴を知る方法として様々な分野で用いられている。ヒストグラムを推定する上で，決めるべき２つのパラメータがある。１つは分割幅（以降，bin幅という）と，もう１つは端点（以降，Anchor Positionという）である。ノンパラメトリック統計理論の中で，平均二乗誤差の漸近論に基づく最適なbin幅が得られており，多くの論文で議論されている。他方，Anchor Positionを決める問題は未解決のままである。

本研究ではAnchor Positionを変化させた時のヒストグラム推定へ与える影響を数値実験で示し，その重要性を明らかにする。そして，モーメント法に準じた方法でAnchor Positionを決める手順を提案し，数値実験により，その有効性を明らかにする。

１．はじめに

ヒストグラムはデータの概要を把握するために古くから利用されている。確率密度関数の推定法としてのヒストグラム推定は統計的手法の中で最も簡単なものの一つである。これはヒストグラムが構築・解釈が容易であり，計算が簡単で高度なグラフィックを必要としないからである。

ヒストグラムの重要なパラメータの１つ目は，bin数（分割数）或いはbin幅をどのように選択するかという問題である。bin幅に関してScottの公式（1979），Freedman and Diaconisの公式（1981），bin数に関してはSturgesの公式（1926），Doaneの公式（1976）などの研究がある。２つ目は，ヒストグラムをど

こから描き始めるか，つまり端点(Anchor Position)を決める問題である。このAnchor Positionを決める問題は,重要な問題でありながらこれまであまり議論されていない。

本研究ではヒストグラム推定に与えるAnchor Positionの影響を調べ,Anchor Positionの選択法を提案し,その数値実験によって有効性を検証する。

２．ヒストグラムの評価基準と安定指数

２．１．ヒストグラムの定義

ヒストグラムとは，データ(観測値)をいくつかの階級に分け，それぞれの階級度数を数えて，グラフ化したものの総称である。グラフの柱状のものをbinといい，横軸にデータの値を取り，その高さはbinの中に入るデータの個数に比例するように決められる。ヒストグラムを密度関数とした場合，高さは,

$$\text{高さ}=\frac{\text{相対度数}}{bin\ \text{幅}},$$

となる。

最も一般的なヒストグラム作成方法である等間隔法とは，ヒストグラムのある固定された区間[a，b]を等間隔に区切って，それをbin幅とする方法である。このbin幅hはヒストグラムを構築する際に重要なパラメータの一つである。ここでn_jはj番目のbinにおけるデータの数，全データ数をnとすると，与えられたビンにおける密度関数$f(x)$のヒストグラム推定は,

$$\hat{f}(x) = \frac{n_j}{nh}, \qquad x \in (b_j, b_{j+1}], \tag{1}$$

となる。ここでb_jは，j番目のbinの左端点とする。

２．２．平均二乗誤差基準

密度関数$f(x)$の推定量$\hat{f}(x)$の評価方法として，平均二乗誤差(*Mean Squared*

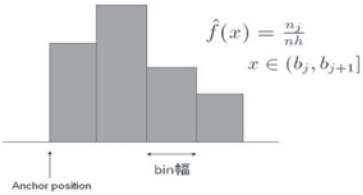

図１．ヒストグラム推定

*Error:*以降*MSE*と記す）がある。これは，

$$\begin{aligned} MSE(\hat{f}(x)) &= E(\hat{f}(x) - f(x))^2 \\ &= Var(\hat{f}(x)) + (E[\hat{f}(x)] - f(x))^2, \end{aligned} \tag{2}$$

と定義される。*MSE*は分散と２乗Biasの和で表されることが分かる。*MSE*は，定点xにおける誤差を表したものである。これに対して，全体的な推定量の誤差の大きさを表すものとして，積分平均二乗誤差（*Integrated Mean S quared Error*：以降*IMSE*と記す）がある。

$$\begin{aligned} IMSE &= \int MSE\{\hat{f}(x;h)\}dx \\ &= IV + ISB, \end{aligned} \tag{3}$$

但し，

$$ISB = \int Bias[\hat{f}(x;h)]^2 dx,$$
$$IV = \int Var[\hat{f}(x;h)]dx,$$

とする。また，漸近的な*IMSE*（*AIMSE*と記す）は*h*に依存することから*AIMS E*（*h*）と表記すると，次式で表される，

$$AIMSE(h) = \frac{1}{nh} + \frac{1}{12}h^2 R(f'), \tag{4}$$

但し，

$$R(f') = \int_{-\infty}^{\infty} f'(x)^2 dx.$$

*AIMSE*を最小とするh^*は，

$$h^* = [6/R(f')]^{1/3} n^{-1/3}, \tag{5}$$

となる。ここで簡便な方法として，もし$f(x)$が平均μ，分散σ^2の正規分布に従うと仮定した場合，$R(f') = 1/(4\sqrt{\pi}\sigma^3)$ となり，⑸式から，

$$h^* = (24\sqrt{\pi}\sigma^3/n)^{1/3} \approx 3.5\sigma n^{-1/3}, \tag{6}$$

となる。Scott（1979）は⑹式において，未知なσを標本分散$\hat{\sigma}^2$によって置き換えることを提案した。

２．３．　bin幅・bin数とAnchor Positionの影響

ヒストグラムを構築する上でbin幅，bin数とAnchor Positionの決定は重要な問題である。bin数に関しては，Sturgesの公式（1926），Doaneの公式（1976）など様々な公式が提案されている。同様にbin幅の決め方は，前述のScottの公式（1979）をはじめ，Freedman and Diaconisの公式（1981），Plug-in法（Wand：1997），Bootstrapped Plug-in法（寒河江，田中，山本：2004）などがある。

しかしながらAnchor Positionに関しては重要な問題でありながら，未解決のままである。ヒストグラムの平滑化を取り扱っているAverage Shifted Histogram（Scott：1985a），Frequency Polygon（Scott：1985），Edge Frequency Polygon（Jones他：1998），Generalized Frequency polygon（寒河江，山本：2000）などAnchor Positionの影響を軽減しているがAnchor Positionの決定法は未解決のままである。

図２－５のヒストグラムは横軸に年齢，縦軸にその年齢の人の割合とした時のあるスポーツチームの年齢分布である。**図２**は年齢分布（bin幅＝１）である。**図３**は，Anchor Positionは**図２**と同じであるが，bin幅hをScottの公式によって決定し，推定したヒストグラムである。**図２**と**図３**と比較すると，**図３**で

は20, 33あたりで本来データがないところで 0 とならない推定値をもつ。またヒストグラムのモード（最頻値）の位置も異なる。図４，５はbin幅をScottの公式で決め，Anchor Positionを17.5，18と動かした時のヒストグラムである。図２－５を比べてみるとモードの位置や凹凸がAnchor Positionの変化とともに大きく変化している様子が分かる。

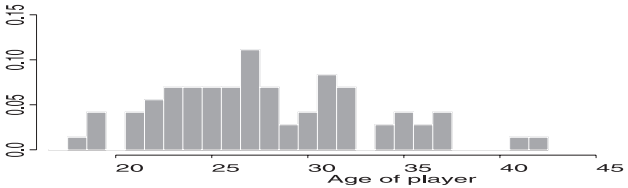

図２．スポーツチームの年齢分布（h= 1 ,Anchor Position=17）

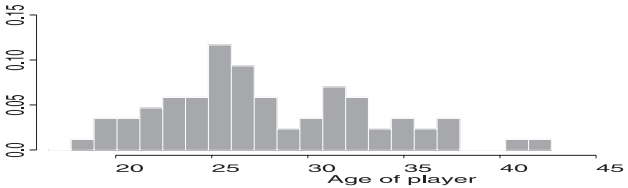

図３．　h=1.189，Anchor Position=17

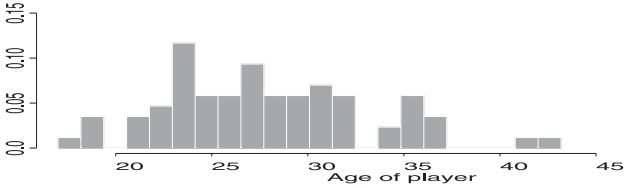

図４．　h=1.189，Anchor Position=17.5

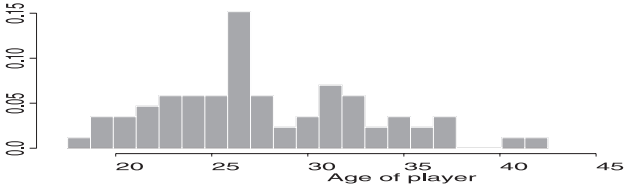

図５．　h=1.189，Anchor Position=18

２．４．　Anchor Positionのヒストグラムへの影響を測る尺度

Simonoff and Udina（1997）は安定指数という指標を導入している。この指数はAnchor Positionの変化によって，推定されたヒストグラムの外観の変化量を測る尺度である。この尺度は，推定されたヒストグラムの凹凸などの類似性を示す指標である。平均二乗誤差は真の分布と推定されたヒストグラムとの誤差を評価しているが，推定したヒストグラムの外観上の変化を捉える指標ではないことに注意する。

ヒストグラム推定に対して，一階導関数f'を隣接するbinごとの変化量と見做し，$\widehat{f'(x)}$は，

$$\widehat{f'(x)} = \frac{n_{j+1} - n_j}{nh^2}. \qquad x \in (b_{j+1} - h/2, b_{j+1} + h/2)$$

と定義する。このとき，$R(f')$の推定は，

$$\widehat{R(f')} = \int \widehat{f'(x)}^2 dx = \sum \widehat{f'(x)}^2 \cdot \mathrm{h} = \frac{1}{n^2 h^3} \sum_{j=0}^{K} (n_{j+1} - n_j)^2 = S,$$

となる。Scott and Terrell（1987）はこの$R(f')$の推定を，⑷式に基づいてbin幅hを決めるためのバイアスクロスバリデーション法に用いた。Simonoff and Udina（1997）は単純にAnchor Positionが移動したときに，ヒストグラムの外観の変化を反映する値として利用した。

S_iをbinの端（bin edge）が$\{b_1 - t,\ t \in [0, \mathrm{h}]\}$のときの$R(f')$の推定値とする。Simonoffは簡単のために，$S_i = S_{b1 - ih/T}$，$i = 1, \cdots, T$と定義した。さらに$S_i$の可変性（ヒストグラムの不安定性）をもっているかどうか判断するために，Gini係数（Marshall and Olkin,1979）を利用した。$q_0 = 0$とし，$i = 1, \cdots, T$に対して，

$$q_i = \sum_{j=1}^{i} S_{(j)} / \sum_{j=1}^{T} S_j,$$

と定義する。ここで$S_{(j)}$はj番目の順序統計量である。そして座標$(i/T, q_i)$，$i = 0, \cdots, T$が描く曲線をローレンツ曲線と言う。

Gini係数の定義から，

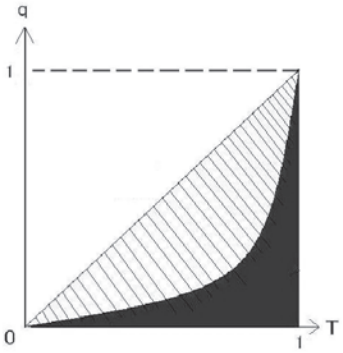

図６．安定指数Gの定義

$$Gini\text{ 係数 } = \text{ データから描かれた図６斜線部の面積} \times 2,$$

ここで，データから描かれる図６斜線部の面積を考える。この面積は対角線とローレンツ曲線に挟まれている部分の面積で，台形の集合と考えることが出来る。したがって，Gini係数は，

$$Gini\text{ 係数 } = 1 - \sum_{i=1}^{T}(q_i + q_{i-1}) \cdot \frac{1}{T}$$

となる。Simonoffはこの第二項（図６黒く塗りつぶした面積×２）を安定指数Gとした。

$$G = 1 - Gini\text{ 係数} = \sum_{i=1}^{T} \frac{q_i + q_{i+1}}{T}, \tag{7}$$

安定指数Gは領域[0,1]に属し，Anchor Positionの変化によって推定されたヒストグラムの凹凸などの変化が小さければ小さいほど１に近い値を示す。SimonoffはこのGの値が0.85以上ならば，Anchor Positionに依らず，推定したヒストグラムが安定した外観を示しているとした。

このGはbin幅hの関数であり，hの値に対するヒストグラムの外観の変化を表す指標として，ヒストグラムを推定する上でbin幅hの重要性を検証するこ

とができる。また，Anchor Positionを変化させた時のヒストグラムの安定度の指標としても考えることができる。SimonoffはこのGを用いてbin幅, Anchor Positionの重要性を述べているが，Anchor Positionの決定法に関しては触れていない。

２．５．　Anchor Positionの決定法

平均値保存則とは，推定関数の期待値と母平均が等しいことである。つまり真の分布をf，推定関数を$\hat{f}$とすると以下の式を満たす条件である；

$$E_{\hat{f}}(X) = E_f(X).$$

ヒストグラムのAnchor Positionをa, そしてbin幅hによる分割数（bin数）をkとすると，平均値保存則は

$$E_{\hat{f}}(X) = \sum_{i=1}^{k} \int_{a+(i-1)h}^{a+ih} x\hat{f}(x)dx,$$

と定義する。ここで, 真の分布fは未知なので標本平均 $\bar{x} = \sum_{i=1}^{n} x_i/n$ を用いる。よって，

$$\sum_{i=1}^{k} \int_{a+(i-1)h}^{a+ih} x\hat{f}(x)dx = \bar{x},$$

この条件を満たすようにaを求めることが，Anchor Positionの決定法である。データ数n，区間数k，各区間に入っているデータ数n_i，bin幅h，Anchor Positionをaとする。ヒストグラムは各binの区間では一様分布と見ることができるので，平均は，

$$\sum_{i=1}^{k} \frac{n_i}{n}(a + ih - \frac{h}{2}) = \bar{x},$$

と表すことが出来る。ここで $\sum_{i=1}^{k} n_i = n$ より，Anchor Position；aについて整理すると，

$$a = \bar{x} + \frac{h}{2} - \frac{h}{n}\sum_{i=1}^{k} n_i i. \qquad (8)$$

各区間のデータ数n_iと区間数kは，Anchor Positionが変化するとそれぞれのbinの位置が変わるので，同じではない。よって平均値保存則を満たすAnchor Positionを求めるには，以下の手順を行う。

1．データの最小点より小さいある点aを初期値$a^{(1)}$として他のパラメータn_i, kを求める。
2．(8)式に代入し，求めたaを新たなAnchor Position $a^{(2)}$としてn_i，kを求める。
3．停止条件：$|a^{(i)} - a^{(i-1)}| < \varepsilon$（$\varepsilon$は0に近似される非常に小さな値）を満たすまで1－3を繰り返す。

以上のように平均値保存則を満たすAnchor Positionを決定する。

数値実験によると，選択可能なAnchor Positionの中に平均値保存則を満たすものは，ただ一つではなく複数存在することがある。したがって，以下の実験では，三つの異なる初期値を用いた平均値保存則によるAnchor Positionを比較している。

3．Anchor Position決定法の数値実験

3．1．データ数によるAnchor Positionへの影響

標準正規分布からのデータに関してデータ数を変化させた時（N＝10，30，50，100，300，500），Anchor Positionを動かすことによる積分2乗誤差*ISE*の変化を示した図である。**図7**は50回の推定結果についての*ISE*を表示したものである。

各図を比較すると，データ数が少ないほどAnchor Positionの位置に対する*ISE*の値の影響度が大きい。このことから，データの少ない場合には，Anchor Positionの決定はより重要な問題となる。

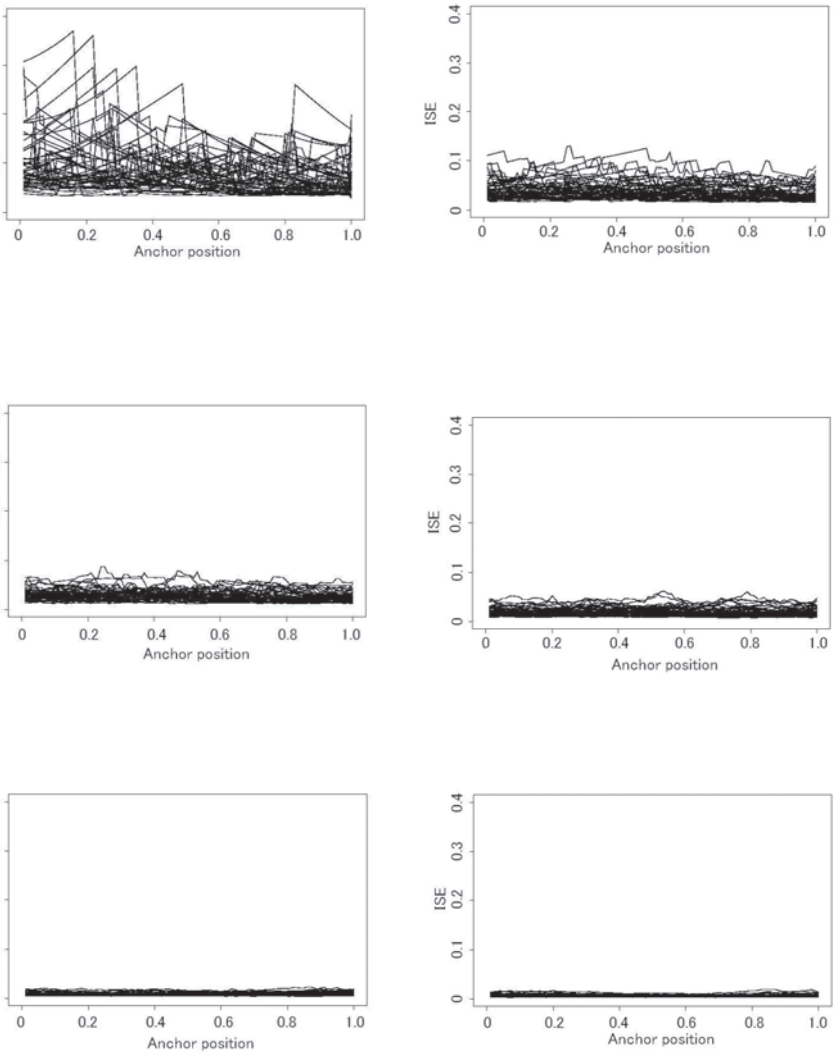

図７．左上図〈*N* = 10〉右上図〈*N* = 30〉左中図〈*N* = 50〉右中図〈*N* = 100〉左下図〈*N* = 300〉右下図〈*N* = 500〉におけるAnchor Positionを動かした時の*ISE*(50回)

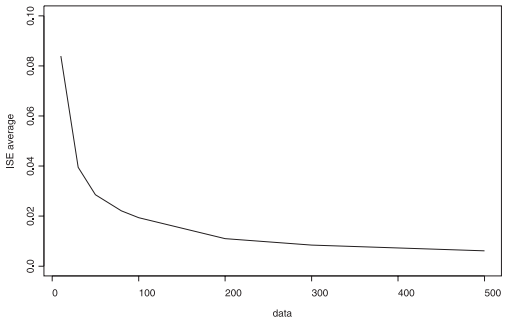

図８．ISE平均

図８は**図７**の*ISE*値の50回の平均値である。データの数が増えれば*ISE*の平均値も小さくなる。データ数が少なければ平均的に*ISE*の値が大きく，Anchor Positionの影響度が大きいことを表している。これはデータ数が増えれば，bin幅の推定値が小さくなりAnchor Positionを移動させても*ISE*の値は安定するためである。

３．２．実験方法

Anchor Positionの数値実験について，Simonoff and Udina（1997）の例を用いて，以下の三つの分布と推定区間について実験を行った;

- 標準正規分布　推定区間 [－4,4],
- 三峰型分布推定区間 [－4,4],
- 歪みのある分布　推定区間 [－4,4].

Anchor Positionとしては，三つの異なる初期値（$min\,(data) - h/2$），（$min\,(data) - h$），（$min\,(data) - 3h/4$）での平均値保存則による決定法と対比のために２つのAnchor Position：$min\,(X_i)$と$\bar{X} - \sqrt{3S}$（但し，Sは標本分散）のと数値比較を行う。$min\,(X_i)$はデータの最小値をAnchor Positionにとることであり，一様分布における区間の最尤推定でもある。又，$\bar{X} - \sqrt{3S}$は一様分布のモーメン

ト推定に基づくAnchor Positionである。以下の手順で実験を行う。

１．各分布に従う乱数を発生させる。
２．発生させたデータに対し，Scottの公式によりbin幅を決定する。
３．Simonoffの安定指数Gを計算する。
４．各方法によって決められたAnchor Positionにおけるヒストグラム推定を行い，誤差ISEを計算する。
５．１～４を1000回繰り返し，最も小さいISEに対応するAnchor Positionの選択法を数え上げる。
６．求められた各決定法のISE値の1000回の平均と分散を求め, 各方法を比較する。

３．２．１．標準正規分布

標準正規分布では，N＝50, 100, 500で数値実験を行った。bin幅hはScottの公式によって決めている。

標準正規分布のヒストグラム推定におけるAnchor Positionの安定指数を比べたのが**表１**である。サンプル数N = 50, 100, 500で比べると，データ数が増えるに従い，安定指数G値が0.85を超える回数が増えている。これは，データ数の増加と最適bin幅が狭くなることで，Anchor Positionの影響が小さくなるからである。このことから小規模データにおいてAnchor Positionの選択がより重要となる。

表１．データ数による安定指数の変化（実験1000回）

安定指数	N＝50	N＝100	N＝500
G＞0.85	875	955	994
G≦0.85	125	45	6

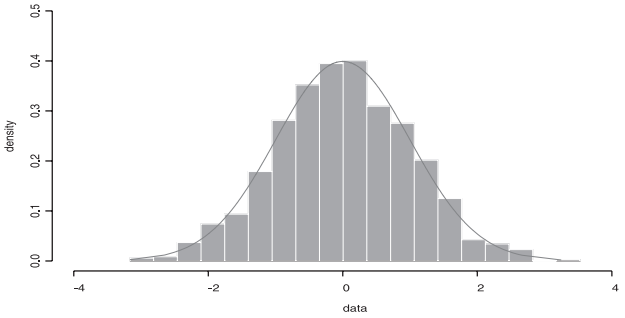

図９．*N*（０，１）

表２はデータ数50における平均値保存則（３つの初期値，$min\,(data) - h/2$, $min\,(data) - 3\,h/4$, $min\,(data) - h$）に基づく三種類のAnchor Positionと$min\,(data)$と$\bar{X} - \sqrt{3S}$をAnchor Positionとする二つの場合について，それぞれのヒストグラム推定値を求め，真の分布との差として*ISE*を求め，*ISE*最小となるAnchor Positionの回数を1000回繰り返して，求めた表である。

表３は，Anchor Position選択法としての三つの初期値での平均値保存則と$min\,(data)$，$\bar{X} - \sqrt{3S}$をAnchor Positionとした時のヒストグラム推定を行い，*ISE*を計算し，1000回の平均を示している。括弧内の数値はその分散である。

表２．データ数50における*ISE*が最小となるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X} - \sqrt{3S}$
$min\,(data) - h/2$	547	361	92
$min\,(data) - 3\,h/4$	580	318	102
$min\,(data) - h$	515	311	174

表３．データ数50におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X} - \sqrt{3S}$
$min\,(data) - h/2$	$0.0275\,(1.721 \times 10^{-4})$	0.0314 (2.182×10^{-4})	0.0367 (2.716×10^{-4})
$min\,(data) - 3\,h/4$	$0.0276\,(2.007 \times 10^{-4})$		
$min\,(data) - h$	$0.0284\,(1.904 \times 10^{-4})$		

表４．データ数100における*ISE*が最小にとなるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	534	419	47
$min\,(data)-3h/4$	559	391	50
$min\,(data)-h$	510	383	107

表５．データ数100におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data-h/2$	$0.0176\,(5.170\times10^{-5})$	0.0185 (6.066×10^{-5})	0.0251 (1.044×10^{-4})
$min\,(data)-3h/4$	$0.0176\,(6.211\times10^{-5})$		
$min\,(data)-h$	$0.0179\,(6.097\times10^{-5})$		

表４，５はデータ数100における1000回の実験における*ISE*が最小となるAnchor Positionの回数をカウントしたものとそのときの*ISE*の平均と分散である。

表６．データ数500における*ISE*が最小にとなるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	493	502	5
$min\,(data)-3h/4$	491	496	13
$min\,(data)-h$	483	502	5

表７．データ数500におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	0.00635 (4.350×10^{-6})	0.00639 (4.621×10^{-6})	0.00961 (7.300×10^{-6})
$min\,(data)-3h/4$	0.00638 (4.588×10^{-6})		
$min\,(data)-h$	0.00638 (4.282×10^{-6})		

表６，７はデータ数500のときの同様な実験の結果である。

N＝50の時，平均値保存則によるAnchor Position決定法は$min\,(data)$と$\bar{X}-\sqrt{3S}$をAnchor Positionとした時と比べて良い推定値が得られた。平均値保存則の初期値に関しては（min（data）－３h/４）が表２，４より最も良い推定値となった。

安定指数Gは，データが少ないため，bin幅の推定値が大きくなり，$G \leq 0.85$になる不安定な外観を示す場合が比較的多く見られた。

N=100の時，ほぼN=50の場合と同様の結果が得られた。平均値保存則における初期値（min（data）－ 3 h/ 4 ）のとき，良い推定値が得られた回数が最も多く，*ISE*の平均値も最も小さかった。ただし平均値保存則と他のAnchor Positionとの間のISE平均の差は小さくなっている。Gに関してはデータが増えているので，$G \leq 0.85$となるヒストグラム推定はN=50に比べて少なくなっている。

N=500の時，平均値保存則の有効性はほとんどなくなっている。*ISE*平均の値もほとんど同じである。これは，データが増えれば，bin幅の推定値hが小さくなり*ISE*に及ぼす影響度が少なくなるためである。Gは$G \leq 0.85$となることがほとんどなく，Anchor Positionを移動させてもヒストグラムの外観が大きく変化することはなくなっている。

標準正規分布において，データが少ない場合に平均値保存則が有効性があることは実証できた。また先に示したとおりデータが大きければAnchor Positionの影響度は小さくなる。このことから，以降の２つの例では，N=100の場合について実験を行う。

３．２．２．三峰型確率分布

三峰型密度関数;

$$f(x) = \frac{1}{3}N(0, 1) + \frac{1}{3}N\left(-2, \left(\frac{1}{3}\right)^2\right) + \frac{1}{3}N\left(2, \left(\frac{1}{3}\right)^2\right),$$

について数値実験を行う。この関数はSimonoff and Udina（1997）の中で使われているものである。

三峰型分布において，初期値を（min（data）－ 3 h/ 4 ）としたものが最も良い効果が得られた。平均値保存則がどの初期値からAnchor Positionを決めても良い結果が得られた。*ISE*の平均値も平均値保存則による結果は他のAnchor Positionと比べてかなり小さな値となった。特に初期値（min（data）－ 3 h/ 4 ）としたときは最も良い推定が得られた。

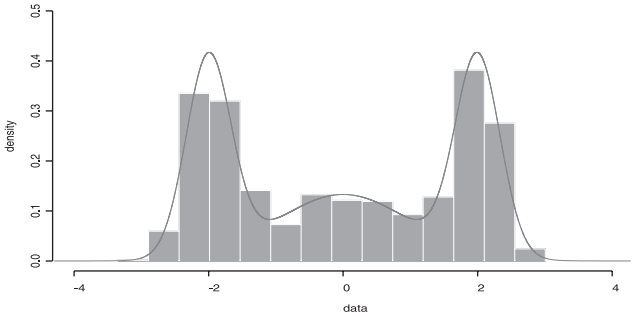

図10．三峰型確率分布

表８．データ数100における*ISE*が最小にとなるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	620	369	11
$min\,(data)-3h/4$	822	171	7
$min\,(data)-h$	731	250	19

表９．データ数100におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	0.0667 (8.732×10^{-4})	0.0734 (3.482×10^{-4})	0.1048 (4.731×10^{-3})
$min\,(data)-3h/4$	0.0517 (2.421×10^{-4})		
$min\,(data)-h$	0.0540 (2.369×10^{-4})		

３．２．３．歪みのある確率分布１

密度関数$f(x)$;

$$f(x)=\sum_{i=0}^{7}\frac{1}{8}N\left(3\{\left(\frac{2}{3}\right)^{i}-1\},\left(\frac{2}{3}\right)^{2i}\right),$$

で与えられる。

歪みのある確率分布はAnchor Position近くの裾で急に立ち上がる関数の例

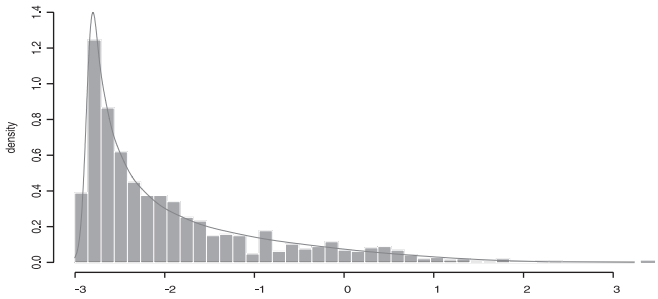

図11．歪みのある確率分布

表10．データ数100における*ISE*が最小にとなるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	11	487	502
$min\,(data)-3\,h/4$	33	480	487
$min\,(data)-h$	20	484	496

表11．データ数100におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	0.3320 (4.935×10^{-3})	0.1954 (2.051×10^{-3})	0.1724 (4.841×10^{-3})
$min\,(data)-3\,h/4$	0.3227 (8.764×10^{-3})		
$min\,(data)-h$	0.2693 (3.319×10^{-3})		

である。他の二つのAnchor Positionと比べて平均値保存則によるAnchor Positionの決定法の精度は著しく低下した。これはこの関数が－３付近にデータが集中しており，データの最小値を選ぶような他の二つの方法に比べ，平均値保存則が－３付近の最小値よりも小さな値をAnchor Positionとするためであると考えられる。このような理由から，平均値保存則による決定法のISEの平均値も他の二つの方法と比べ大きな値となっている。それに対し，

Anchor Positionを最小値として$min\,(data)$や,$\bar{X}-\sqrt{3S}$を選んだ方が良い結果が得られた。このような急激な立ち上がりやデータの集中がAnchor Position付近にある場合の平均値保存則での対処法を次に考える。

３．２．４．歪みのある確率分布２

歪みのある密度関数$f(x)$;

$$f(x)=\sum_{i=0}^{7}\frac{1}{8}N\left(3\{1-\left(\frac{2}{3}\right)^{i}\},\left(\frac{2}{3}\right)^{2i}\right)$$

で与えられる例である。この歪みのある確率分布は３．２．３の分布と左右対称な関数である。この関数のヒストグラム推定を行うことは歪みのある分布に対してAnchor Positionを左端ではなく右端から行うことと同等である。

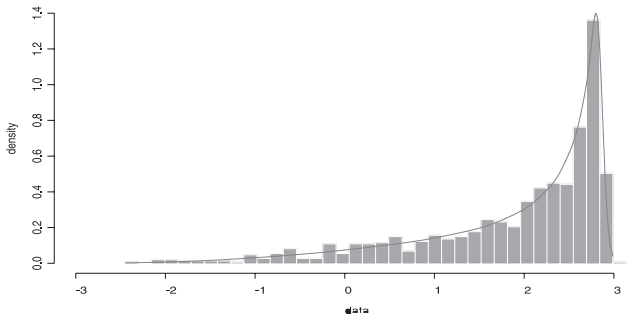

図12．歪みのある確率分布

表12．データ数100における*ISE*が最小にとなるAnchor Positionの回数（実験1000回）

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	601	372	18
$min\,(data)-3h/4$	544	393	63
$min\,(data)-h$	598	297	105

表13. データ数100におけるAnchor Positionによるヒストグラム推定の*ISE*

	平均値保存則	$min\,(data)$	$\bar{X}-\sqrt{3S}$
$min\,(data)-h/2$	0.1594 (1.247×10^{-3})	0.1809 (2.772×10^{-3})	0.2040 (7.297×10^{-3})
$min\,(data)-3h/4$	0.1598 (1.357×10^{-3})		
$min\,(data)-h$	0.1575 (1.150×10^{-3})		

急激な立ち上がりを示す付近でAnchor Positionの選択を避け，この場合，最大値をAnchor Positionとして用いると，平均値保存則による Anchor Position決定法の推定値が最も良かった。平均値保存則による*ISE*の平均値も前述の例と比べて，大きく改善され，他のAnchor Positionと比べて平均値保存則の*ISE*が最も小さな値となった。

このことから歪みのある関数（大きく左に偏っている）の場合には，Anchor Positionをデータの最小値を基準とするのではなく，データの最大値付近を基準に，平均値保存則でAnchor Positionを決定することでISEの値を小さくするようなヒストグラム推定を行うことが出来る。

4．結　　論

ヒストグラム推定において，Anchor Positionを決める問題はSimonoff（1995）によって注意が喚起され，このAnchor Positionの影響によって推定値の外観が大きく変化することが指摘された。又，彼はその変化の度合いを測る安定指数を導入した。

本稿ではSimonoffでは，未解決のまま残されたAnchor Positionの選択法として，ヒストグラムの推定量が平均値（標本平均）を保存するようにAnchor Positionを決める方法を提案し，いくつかの分布について数値実験で有効性を示した。

この方法は，データの集中した所にAnchor Positionを選択する場合，又，今回取り上げなかったが打ち切りデータのような場合，種々のbin幅推定の方

法とAnchor Position選択法共に必ずしも良い結果は与えないことを注意する必要があろう。

参考文献

Doane, D. P.(1976): Aesthetic Frequency Classifications, *The American S tatistician*, Vol.30, No.4, 184－183.

Freedman, D. and Diaconis, P.(1981): On The Histogram as a Density Estimator: L_2 Theory, Zeitschrift fuer wahrsheinlichkeitstheorie und Verwanste Gebeite, Vol.57, 453－476.

Jones, M. C., Samiuddin, M., Al-harbey, A.H. and Maatouk, T.A.H. (1998): The Edge Frequency Polygon, *Biometrika*, Vol.85, No.1, 235－239.

Marshall, A. W. and Olkin, I.(1979): Inequalities: theory of majorization and its applications, Academic Press.

寒河江雅彦，山本けい子(2000): On a generalized class of frequency polygon, Proceedings on 第２回研究集会：「ノンパラメトリック・ファンクショナル推定の理論と応用」，129－144.

寒河江雅彦，田中真寛，山本けい子(2004):ヒストグラムの分割幅と分割数, Proceedings on 第６回研究集会：「ノンパラメトリック・セミパラメトリック法を用いた統計解析理論とその学際的応用」，245－271.

Scott, D. W.(1979): On Optimal and Data-Based Histograms, *Biometrika*, Vol.66，605－610.

Scott, D. W.(1985): Frequency Polygons, *Journal o f the American S tatistical Association,* Vol.80, 348－354.

Scott, D. W.(1985): Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions, *Ann.S tatist*, Vol.13, 1024－1040.

Scott, D. W. and Terrell, G. R.(1987): Biased and unbiased cross-validation in density estimation, *Journal o f the American S tatistical Association*, Vol.82, 1131－1146.

Scott, D. W.(1992): Multivariate Density Estimation(Treory, Practice, and Visualization), New York: Jone Wiley.

Simonoff, J. S.(1995): THE ANCHOR POSITION OF HISTOGRAMS AND FREQUENCY POLYGONS：QUANTITATIVE AND QUALITATIVE SMOOTHING, *Comm. Statist. Simuulation* Comput., Vol.24, 691－710.

Simonoff, J. S. and Udina, F.(1997): Measuring the stabilityof histogram appearance when the anchor position is changed, *Computational Statistics & Data Analysisis*, Vol.23(1997), 335－353.

Sturges, H. A.(1926): The Choice of a Class Interval, *Journal o f the American S tatistical Association,* Vol.21, 65－66.

Wand, M. P.（1997）: Data-Based Choice of Histogram Bin Width, *The American Statistician*, Vol.51, No.1, 59－64.

Wand, M. P. and Jones, M. C.（1995）, Kernel Smoothing, London: Chapman and Hall.

著者連絡先：〒 920－1192　石川県金沢市角間町
金沢大学人間社会研究域経済学経営学系
寒河江雅彦
e-mail: sagae.masahiko@gmail.com

A Selection Method of Anchor Position on Histogram

Masahiko SAGAE[1,*] and Toshihiro YAMAMOTO[2]

[1] Faculty of Engineering, Gifu University

[2] Hitachi, Ltd.

Abstract

Histogram is most commonly used in statistical tools. It is easy to understand and visualize the datas regardless of statistical knowledge. We need to determine two parameters for estimating the histogram. The one is binwidth (or number of bin) and the other is a bin edge (we call it Anchor Position through this paper). The binwidth selection method is proposed by verious authors. On the other, the selection of Anchor Position is undeveloped, asfarasl know. In this paper, it is shown that the change of Anchor Position affects histogram estimation in numerical examples. And it is pointed out that a selection of Anchor Position is important for estimating histogram. We propose the method for selecting Anchor Position based on the method of moments. It is shown that the method works well in some numerical examples.

Key words：Histogram, Anchor Position, Nonparametric Density Estimation, Binning

*Corresponding author

E-mail address: sagae.masahiko@gmail.com (Masahiko SAGAE)